Automating Anomaly Detection

Dongyu Zheng

Search Logging & Metrics, Yelp Inc.

d28zheng@edu.uwaterloo.ca

1 Introduction

The Search Logging & Metrics (SLAM) team at Yelp tracks numerous metrics critical to the company's success. With hundreds of A/B tests actively running and ongoing risks of system failures, it is vital to detect anomalous behavior as soon as possible. With over a thousand metrics, it is infeasible to manually inspect them on a regular basis. The Search Metrics Anomaly Detector (SMAD) is proposed as a tool to automatically detect and alert on such failures. First, this report presents several methods of detecting outliers in highly seasonal, non-stationary, univariate streaming data and compare their performance. Then, methods of detecting anomalies in A/B test metrics are shown, and an overview of SMAD's implementation is given.

One traditional approach towards monitoring metrics for anomalies is through visual inspection on some plotting dashboard. However, not only does manual inspection fail to scale easily, it is also prone to human error. In SLAM's case, there are hundreds of metrics multiplied by hundreds of A/B experiments and their groups, and again multiplied by other dimensions (e.g., Android vs. iOS)—clearly, human monitoring is a poor approach. In the case of A/B testing, failures may only exist in specific groups and are unnoticeable without isolation. Moreover, their implementers often only pay attention to whether the metric they are targeting improves and neglect to ensure other metrics are not adversely affected. For example, increasing the number of ads displayed on the search results page may increase ad revenue, but may also damage the click-through rate metric if the ads displayed are not very relevant.

SMAD is proposed as a tool to automatically detect anomalies in global metrics (metrics where users are not part of any A/B test, i.e., they are in the status quo), as well as detect statistically significant differences between the status quo and experimental cohorts in active A/B tests. Upon detection, an email is to be sent out to relevant parties. On a high level, SMAD ingests data on a nightly basis from Amazon Redshift, performs global and experiment metric analysis, and sends out emails if applicable.

2 The Data

Section removed for confidentiality purposes. This section previously contained information about important/relevant database schemas.

3 What is an Anomaly?

3.1 Outliers in Global Metrics

An outlier in global metrics is characterized by a sudden, unpredicted increase or decrease in value. Outliers often indicate system failures, including logging failures, or failing/poor performing A/B experiments. Values after a concept shift should initially be labeled as outliers, but should not be labeled as outliers after some time.

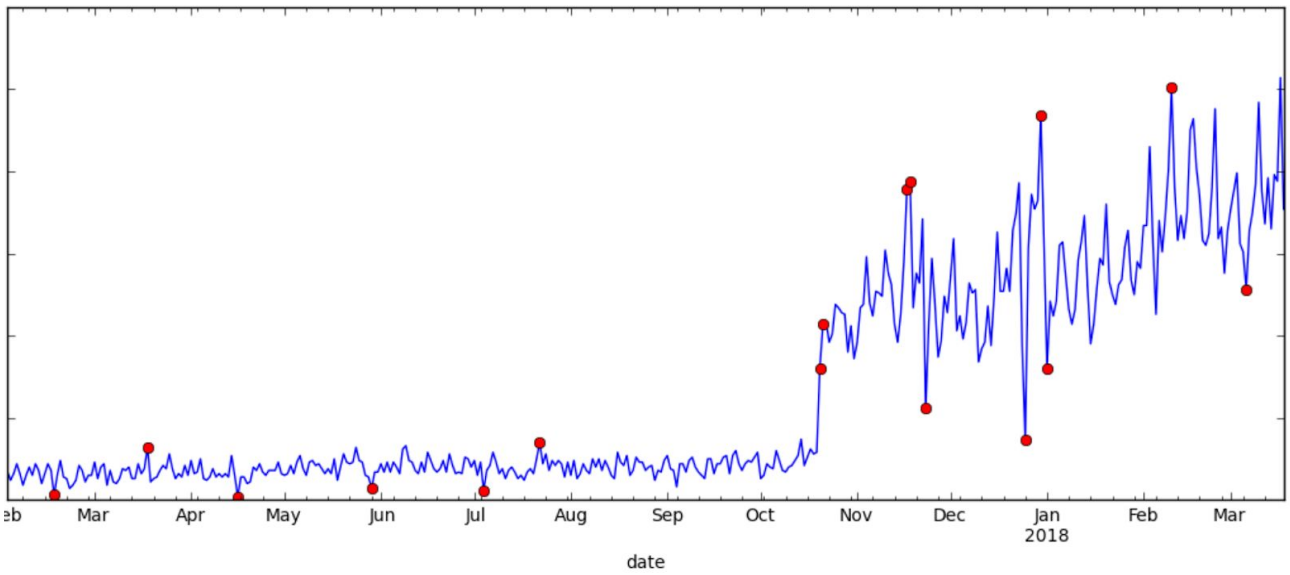

Figure 1: Sample plot of outliers in global metrics

In Figure 1, the blue line is a metric and red dots are outliers. Notice that a concept shift occurs in mid-October 2017, and the values are initially categorized as outliers but are not afterward.

3.2 Statistical Differences in Experiment Metrics

In A/B testing, it is possible for failures to only affect certain experiment cohorts—a bug in one of the cohorts, for example. For small testing groups, these failures are unnoticeable on a global scale. Furthermore, experiments affect much more than just one metric: an experiment that increases some

metrics may decrease others. When analyzing experiment metrics, statistically significant differences between experimental cohorts and the status quo should be reported as anomalies. Generally, these differences can be visualized as "parallel lines" when plotting each cohort.

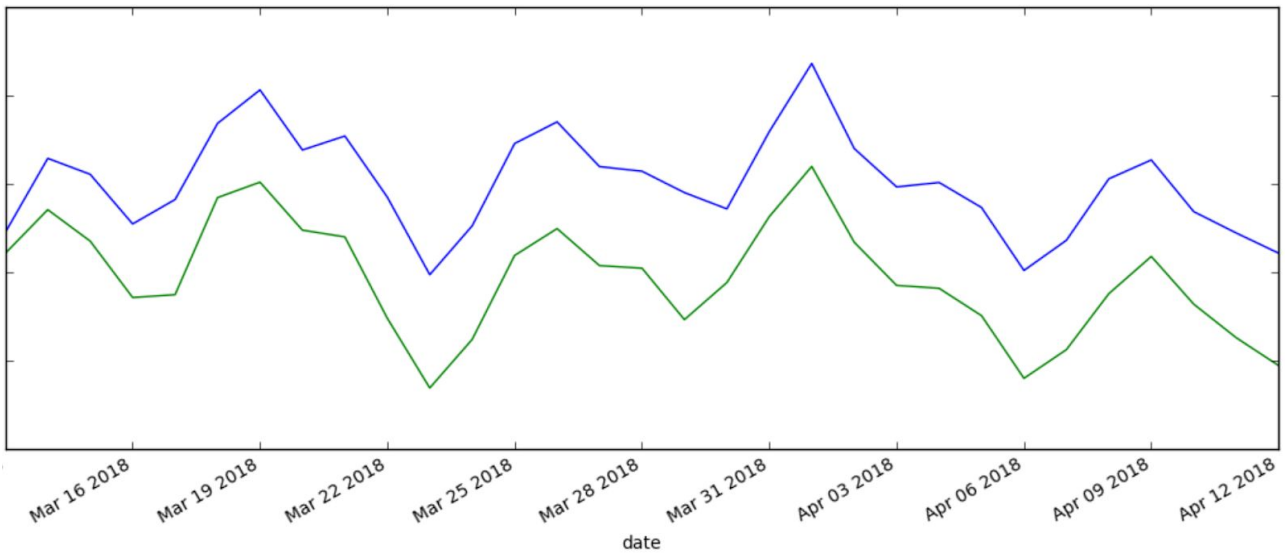

Figure 2: Sample plot of two experiment cohorts which are statistically different

Figure 2 shows the plot of two experiment cohorts (blue is the status quo and green is an experimental cohort, for example). Only statistically significant differences should be reported to avoid a high recall and low precision situation.

4 Detecting Outliers in Global Metrics

In this section, several methods for detecting outliers in highly seasonal, non-stationary, univariate streaming data are presented. Methods are evaluated qualitatively due to a lack of labeled data in SLAM, making scoring infeasible.

4.1 Derivatives

One of the simplest ways to detect anomalies in time series data is to look at its derivatives. If a derivative is an outlier compared to past derivatives, then an anomaly is likely to have occurred. This method can be reasoned intuitively:

- A "flatline" occurs: derivative is too low compared to past derivatives
- A "spike" occurs: derivative is too high compared to past derivatives

$$d_t = \frac{y_t - y_{t-p}}{p} \tag{1}$$

$$[Q_1 - k(Q_3 - Q_1),\ Q_3 + k(Q_3 - Q_1)] \tag{2}$$

For each day, t, the derivative, d_t, is approximated using (1), where y_t is the metric value at t, and p is some period of time, in days. This notation will be used throughout the report. Outliers can then be looked for using Tukey's fences [7], where derivatives outside of the range (2) are flagged as outliers—Q_1 and Q_3 are the first and third quartiles, respectively; and k is some multiplier (usually 1.5). Non-stationary derivatives can be accounted for by calculating the quartiles using an exponentially weighted model, or, more simply, with a rolling window. A rolling window was implemented for the sake of simplicity.

Through experimentation, it was found that a period of $p=5$ worked the best, however, this made no sense intuitively as 5 days is not a real-world seasonality, while 7 days (one week) is. Additionally, it was found that a window size too small or too large resulted in too many false positives. Using a window size of 30 days and $k=1.5$, Figure 3 and Figure 4 are obtained when tested on two sample metrics.

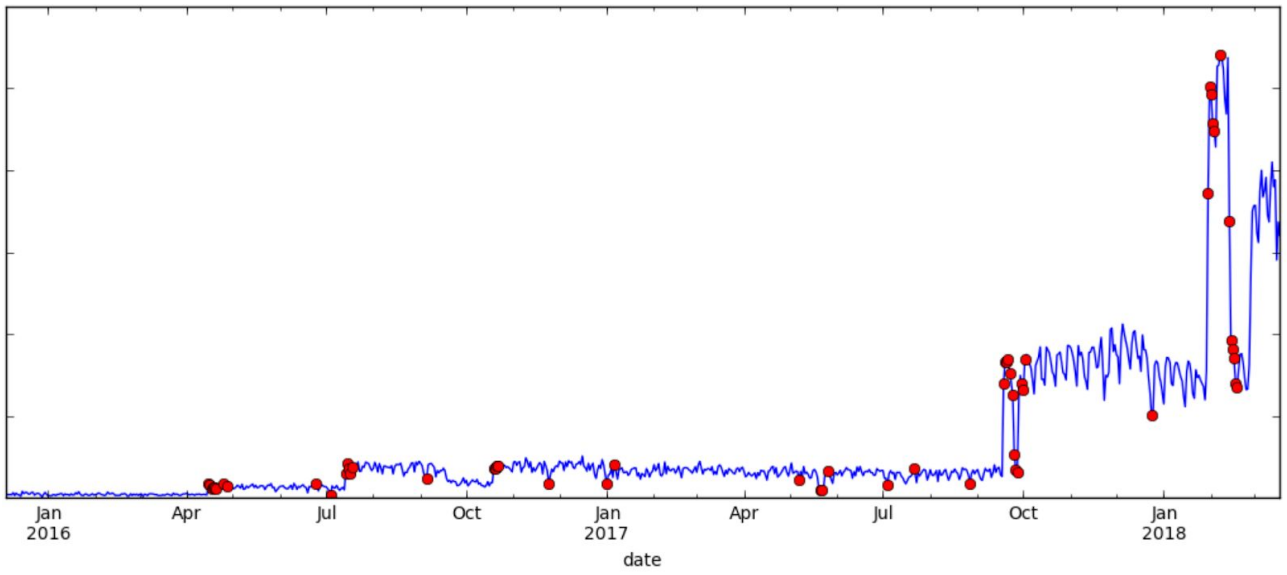

Figure 3: Tukey derivative model on metric A with $k=1.5$

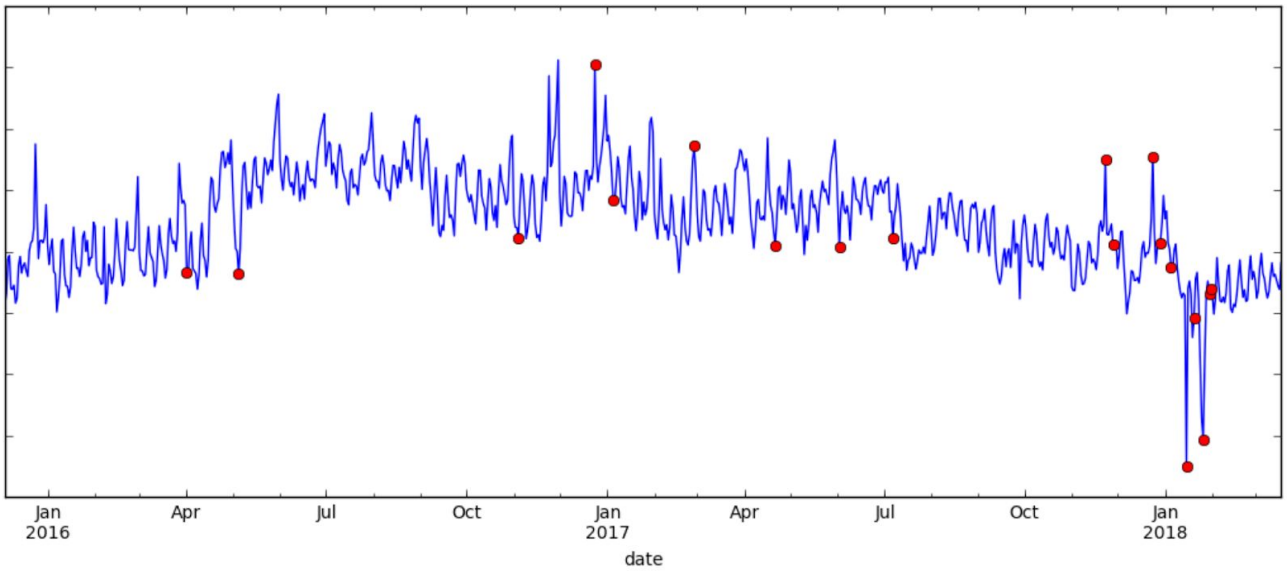

Figure 4: Tukey derivative model on metric B with *k=1.5*

In Figure 3, it is observed that this model performs poorly after concept shifts and generally has too many false positives. Increasing the multiplier to *k=3.0* yields Figures 5 and 6.

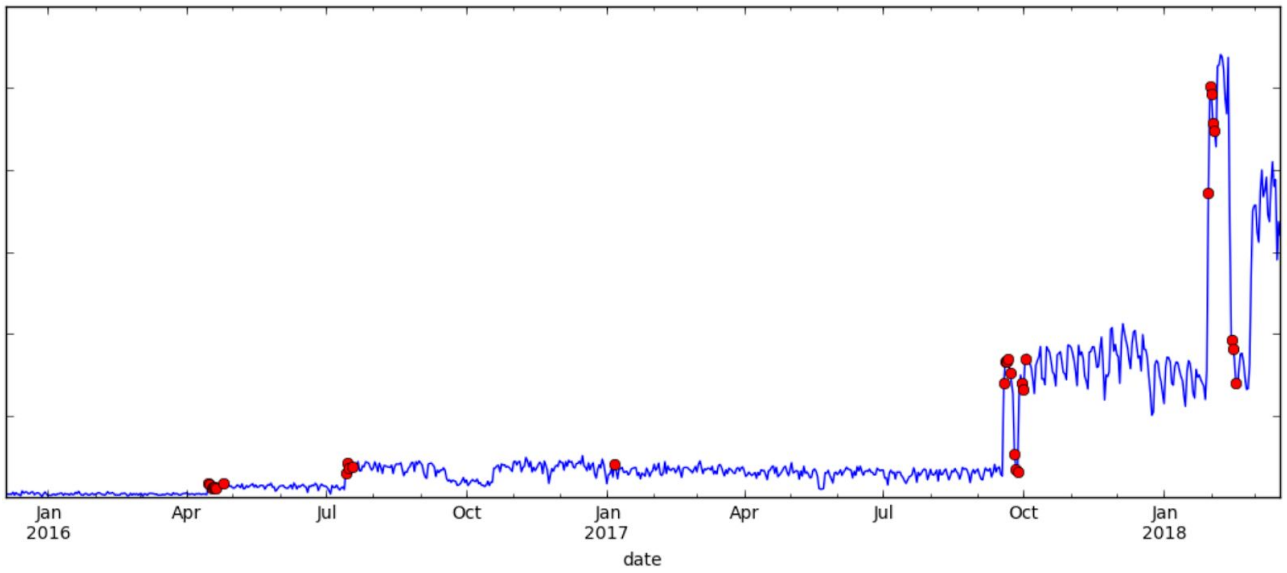

Figure 5: Tukey derivative model on metric A with *k=3.0*

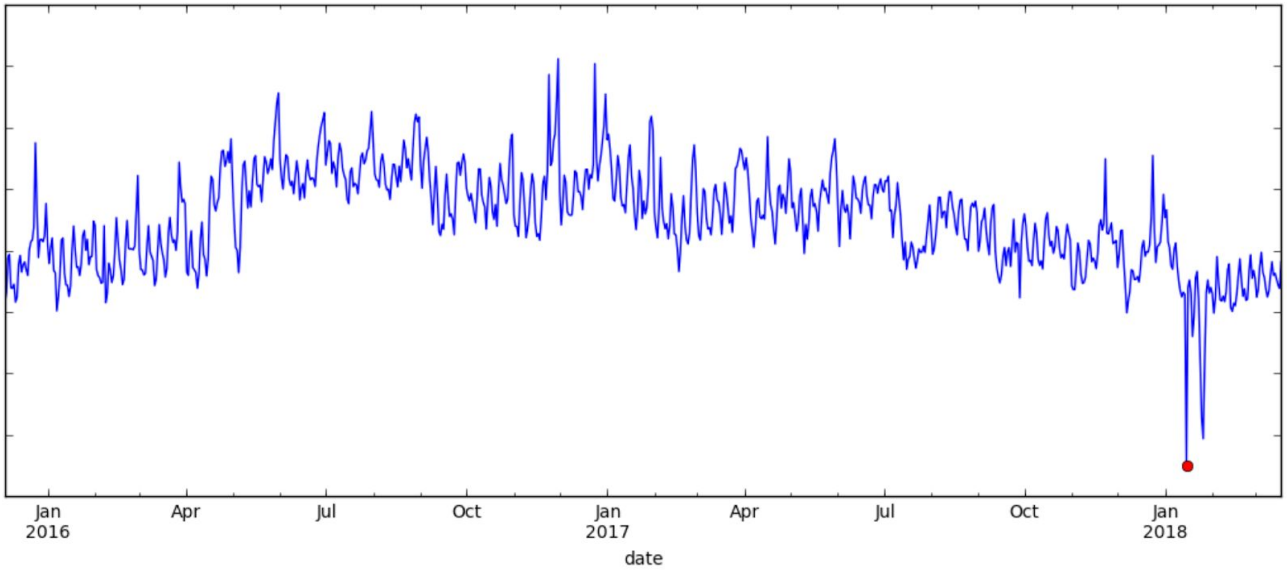

Figure 6: Tukey derivative model on metric B with *k=3.0*

Here, we still observe too many false positives after concept shifts. More importantly, there is an unacceptable amount of false negatives in metric B. Similar observations are made when this model is tested on other samples. In general, it is observed that this model works well for stationary data, but performs poorly on non-stationary ones. Note that a period of *p=1* is the same as using the previous day's value as a prediction and performing a Tukey's fence analysis on the residuals.

4.2 Moving Average

Another simple way of detecting anomalies is to use moving averages to make one-step predictions. The same Tukey's fence analysis, as shown in Section 4.1, can then be performed on the residuals. This moving average model should not be confused with the statistical moving-average model. The cumulative moving average recurrence is presented in (3) below, derived in (1) to (4) from [9].

$$\bar{y}_t = \bar{y}_{t-1} + \frac{y_t - \bar{y}_{t-1}}{t}, \qquad \bar{y}_0 = y_0 \tag{3}$$

$$\hat{y}_t = \bar{y}_{t-1} \tag{4}$$

$$e_t = y_t - \hat{y}_t \tag{5}$$

Calculating the moving average with (3), the predictions are then simply as presented in (4), and the residuals are computed in (5).

Again, using a rolling window approach to deal with non-stationary data, Tukey's fence is used to look for outliers in the residuals. Testing on the same data from Section 4.1 with *k=1.5* and a window size of 30, Figures 7 and 8 are produced.

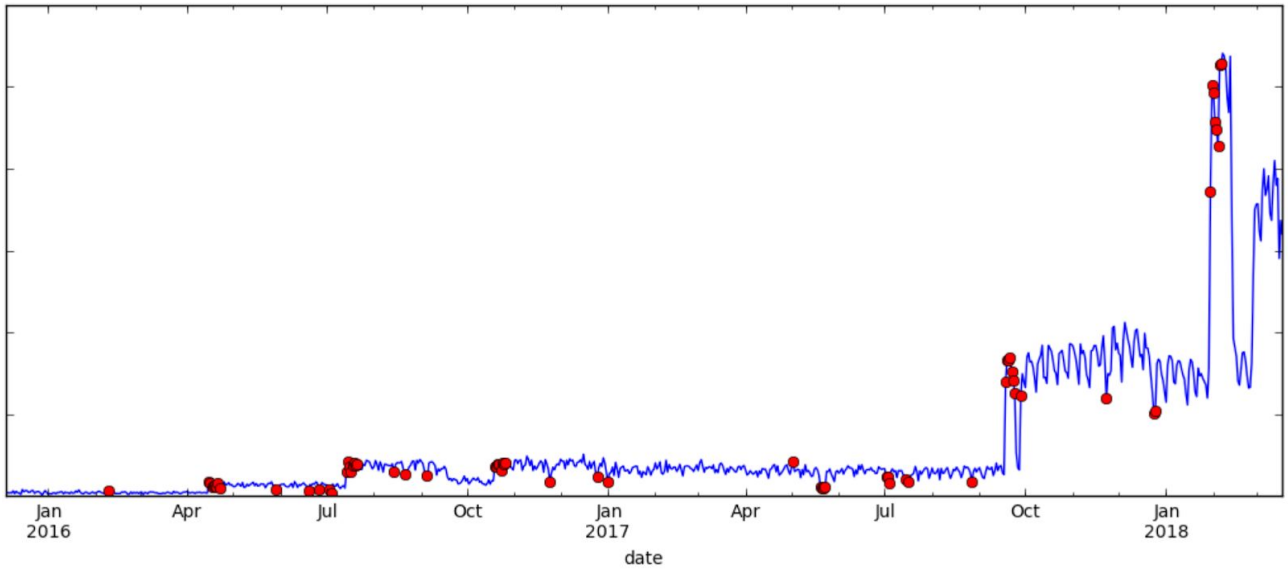

Figure 7: Tukey moving average model on metric A

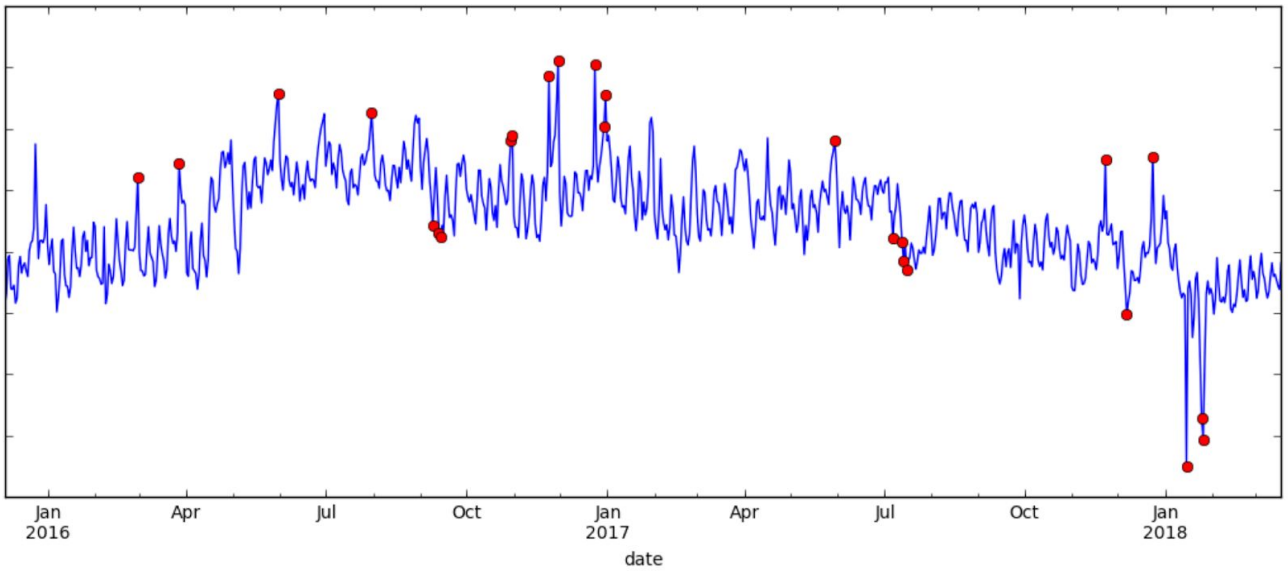

Figure 8: Tukey moving average model on metric B

Similar to the Tukey derivative model, this model does not work well after concept shifts. Increasing the multiplier to *k=3.0* gives very similar results to Figures 5 and 6: there are still a lot of false positives after the concept shift in metric A, and there are too many false negatives in metric B.

This model's poor performance after concept shifts can be attributed to the fact that equal weights are given to all historical values when computing the moving average. This causes predictions

following a concept shift to be very inaccurate for a long duration of time, resulting in false positives. Additionally, this model is not robust to seasonality in the data.

4.4 Exponentially Weighted Moving Average

Following Section 4.3, the exponentially weighted moving average (EWMA) model provides an excellent improvement upon the simple moving average model's issue with concept shifts. Equation 6 below presents the EWMA recurrence derived in (122) to (124) from [9]. Using (6), then (4) and (5), a windowed Tukey's fence analysis can be made on the residuals.

$$\bar{y}_t = \alpha \bar{y}_{t-1} + (1-\alpha) y_t \,, \qquad \bar{y}_0 = y_0 \tag{6}$$

Alpha can be selected as suggested in [1] and [2], or can also be tuned such that the mean squared error of the predictions is minimized. Through experimentation, $\alpha=0.6$ proved to have a fairly low mean squared error on the same sample data as before. Using a window size of 30 and $k=2.0$, excellent results are obtained, shown in Figure 9 and Figure 10.

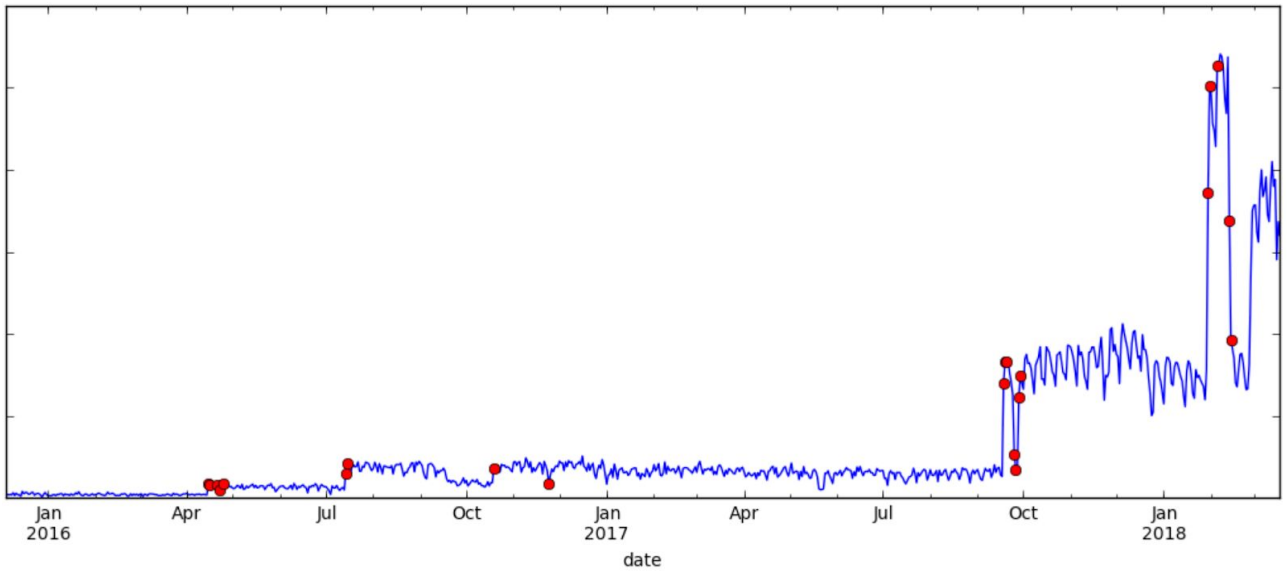

Figure 9: Tukey EWMA model on metric A

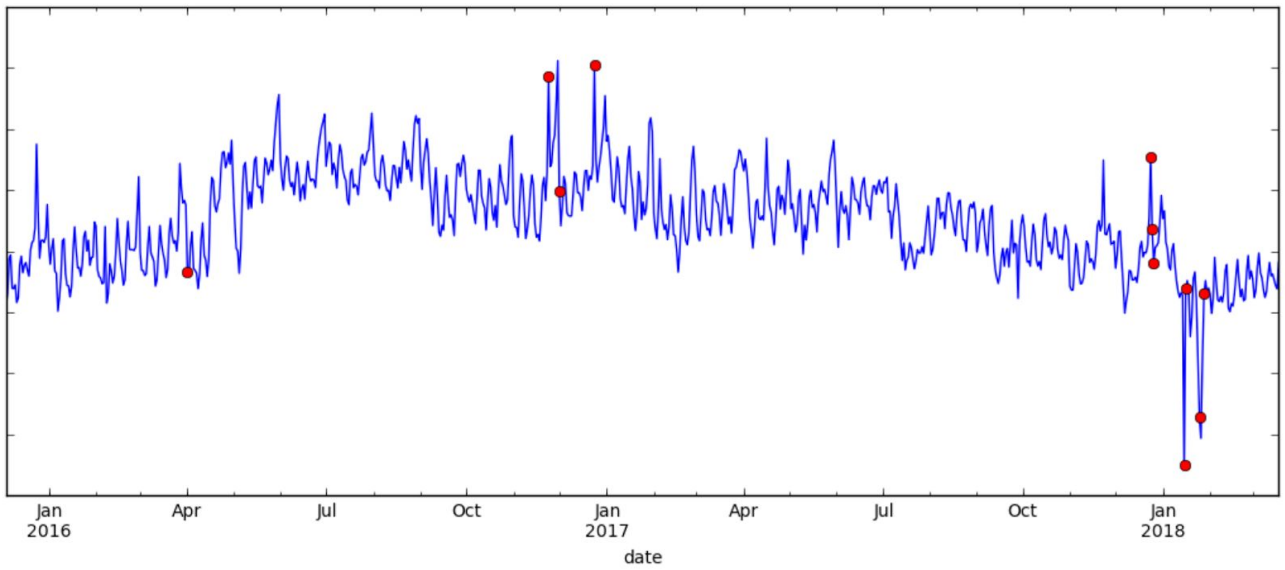

Figure 10: Tukey EWMA model on metric B

With Shewhart charts, a mean and standard deviation analysis can be made on the residuals, under the assumption that the residuals are normally distributed. This is the first stage of the model proposed in [4]. The exponentially weighted moving variance of the residuals can be calculated as shown in (7), derived in (140) to (143) from [9], where $\bar{e}$ is the EWMA of e with $\alpha=\lambda$. The exponentially weighted moving standard deviation of the residuals is then simply (8).

$$S_t = (1-\lambda)(S_{t-1} + \lambda(e_t - \bar{e}_{t-1})^2), \qquad S_0 = 0 \tag{7}$$

$$\sigma_t = \sqrt{S_t} \tag{8}$$

$$LCL_t = \bar{e}_{t-1} - k\sigma_{t-1} \tag{9}$$

$$UCL_t = \bar{e}_{t-1} + k\sigma_{t-1} \tag{10}$$

The lower control limit (LCL) and upper control limit (UCL) are then calculated as presented in (9) and (10), respectively, where k is some multiplier (usually 3.0). A day's value is flagged as an outlier if that day's residual is outside of [*LCL*, *UCL*]. Using $\alpha=0.6$, $\lambda=0.05$, and $k=3.0$, excellent results are obtained in Figures 11 and 12. These results are better than those obtained with a Tukey's fence analysis.

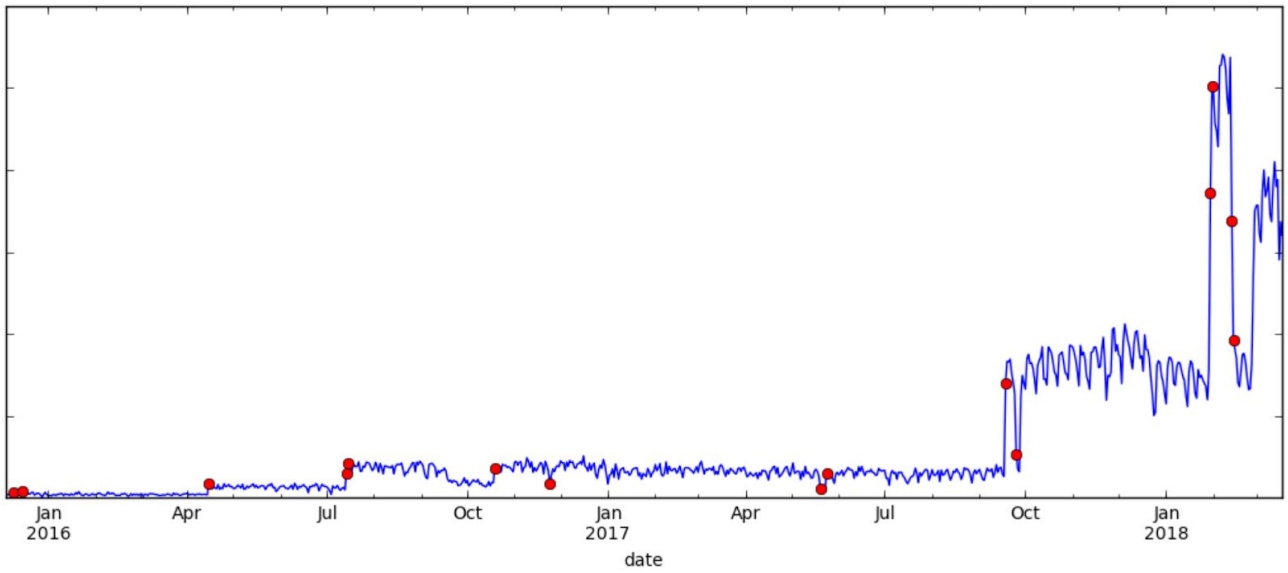

Figure 11: Shewhart EWMA model on metric A

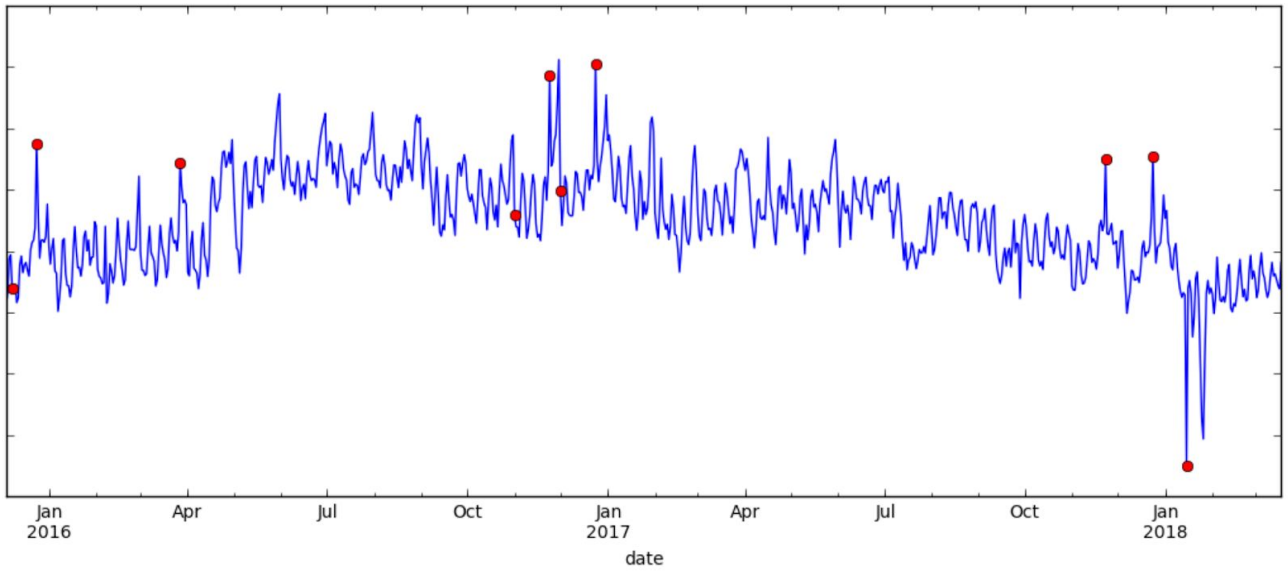

Figure 12: Shewhart EWMA model on metric B

For SLAM's data, the Shewhart EWMA model performs well compared to the Tukey EWMA model and is much better compared to the Tukey moving average model and the Tukey derivative model.

4.5 Weighted Sum EWMA

Observing the fact that real-world seasonalities are in periods of weeks, months, and years, an improvement can then be made to the Shewhart EWMA model, where the predicted $\hat{y}$ is a weighted sum of EWMAs of values from prior days, weeks, months, and years. These EWMAs can be

computed using (11), where p is some period of days (e.g., $p=1$ for day or $p=7$ for week), and α_p is the α used in the EWMA for this specific period—$\bar{y}_{p,\,t}$ is read as the EWMA for period p on day t.

$$\bar{y}_{p,\,t} = \alpha_p \bar{y}_{p,\,t-p} + (1-\alpha_p) y_t\,, \qquad \bar{y}_{p,\,0} = y_0 \tag{11}$$

$$\widehat{y}_t = w_{p_1} \bar{y}_{p_1,\,t-1} + w_{p_2} \bar{y}_{p_2,\,t-1} + \ldots + w_{p_n} \bar{y}_{p_n,\,t-1} \tag{12}$$

Equation 12 shows the weighted sum of the EWMAs using weights w_{pi}. The same Shewhart chart analysis as presented in Section 4.4 is then applied. Furthermore, an improvement to (11) can be made in software implementations where past values are intelligently indexed; e.g., in Python, `dateutil.relativedelta` can be used to difference months, rather than simply using date subtractions of 30 days, which leads to inaccuracies. There are many hyperparameters to optimize for this model: the decays for the EWMAs and the weights for the summation. Grid search and random search using distributed software such as MapReduce and Spark may be suitable for minimizing the mean squared error in predictions. Using randomly searched hyperparameters, $\lambda=0.05$ and $k=3.0$, Figure 13 and Figure 14 are obtained.

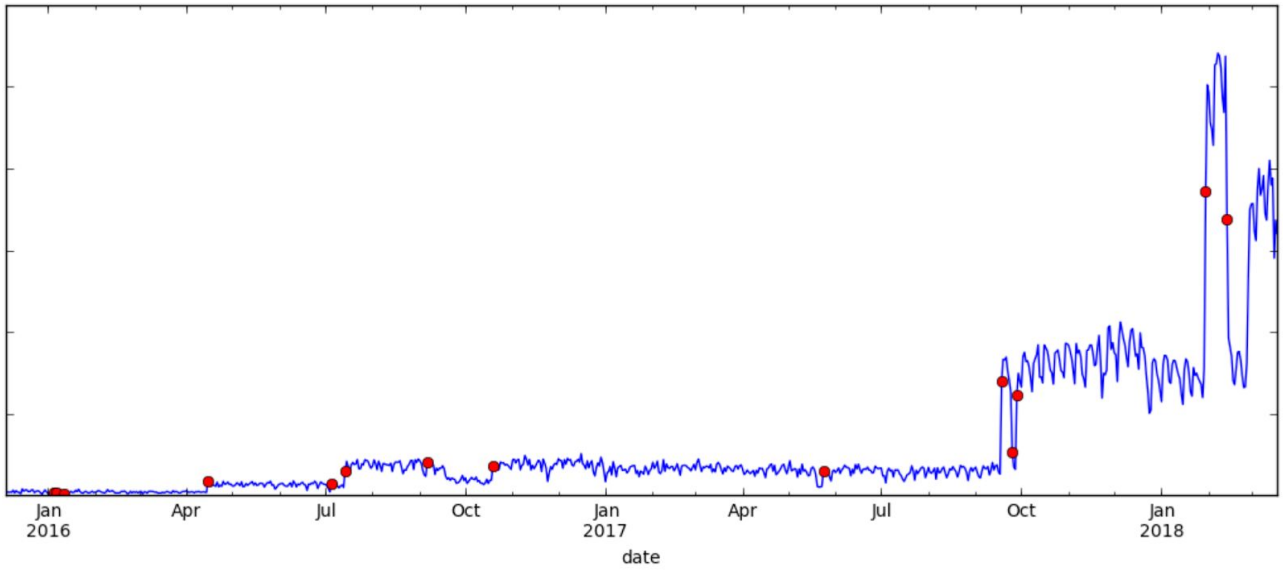

Figure 13: Shewhart weighted sum EWMA model on metric A

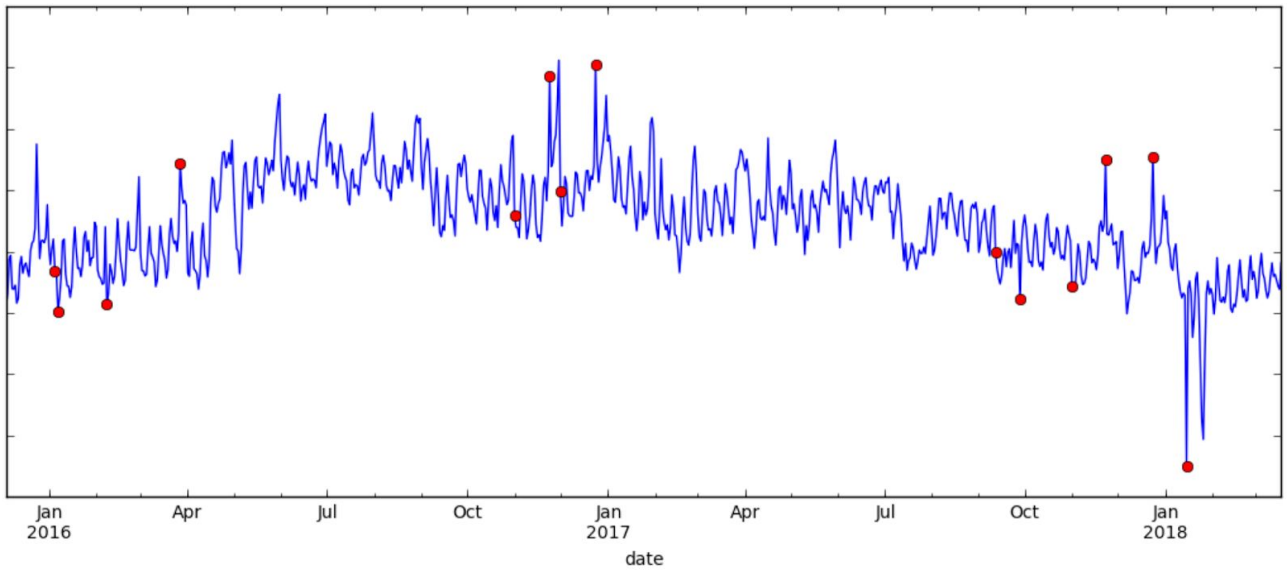

Figure 14: Shewhart weighted sum EWMA model on metric B

Analysis of metric A showed very similar results to that of the Shewhart EWMA model, and analysis of metric B was able to find more anomalies. For SLAM's data, in general, it has been found that the Shewhart weighted sum EWMA model outperforms the other models presented in this report.

5 Detecting Differences in Experiment Metrics

When looking at A/B tests, SMAD aims to identify cohorts which have statistically different performance from the status quo on a certain day.

5.1 Categorical Metrics

In analyzing categorical metrics, the G-test is used to calculate p-values for each metric, client type (e.g., Android vs. iOS), experimental cohort, experiment, and day, under the null hypothesis that the experimental cohort performs the same as the status quo for that specific dimension. The G statistic follows a chi-square distribution and is presented in (13), where E_i are the expected frequencies and O_i are the observed frequencies.

$$G = 2 \sum_{i=1}^{m} O_i \, ln(\frac{O_i}{E_i}) \tag{13}$$

The G-test is used rather than Pearson's chi-square test because of recent recommendations [6]; William's correction is also used [8]. Fisher's exact test is not required because all metrics analyzed have sufficiently large samples.

5.2 Continuous Metrics

Welch's *t*-test is used to calculate p-values for continuous metrics, under the same null hypothesis presented in Section 5.1. Nonparametric tests such as the Mann–Whitney U test are not used because they are computationally infeasible for the input data. Instead, an assumption is made such that cohort populations follow a normal distribution, enabling the usage of Welch's *t*-test.

5.3 False Positive Rate Control

Since so many comparisons were being made, the number of false positives was unacceptably high. Using the two-stage Benjamini, Krieger, & Yekutieli false detection rate procedure [10], the false positive rate was reduced, but experiments such as Figure 15 and Figure 16 were being flagged.

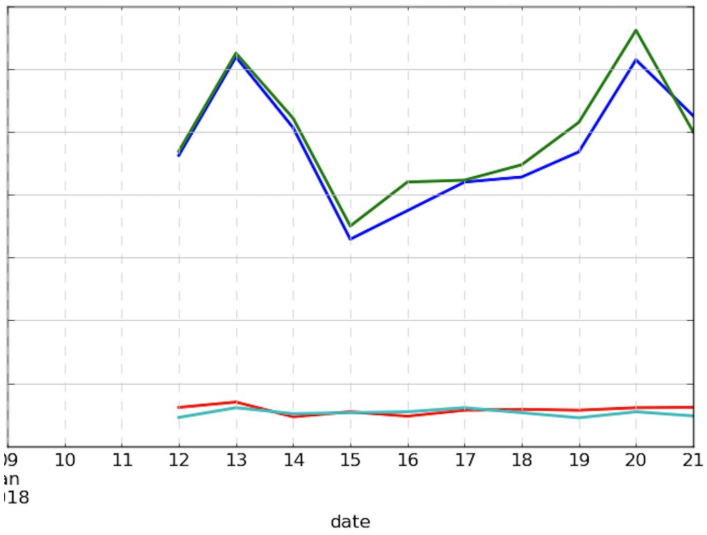

Figure 15: Two cohorts (green, blue) that differ, but have their metric values "cross"

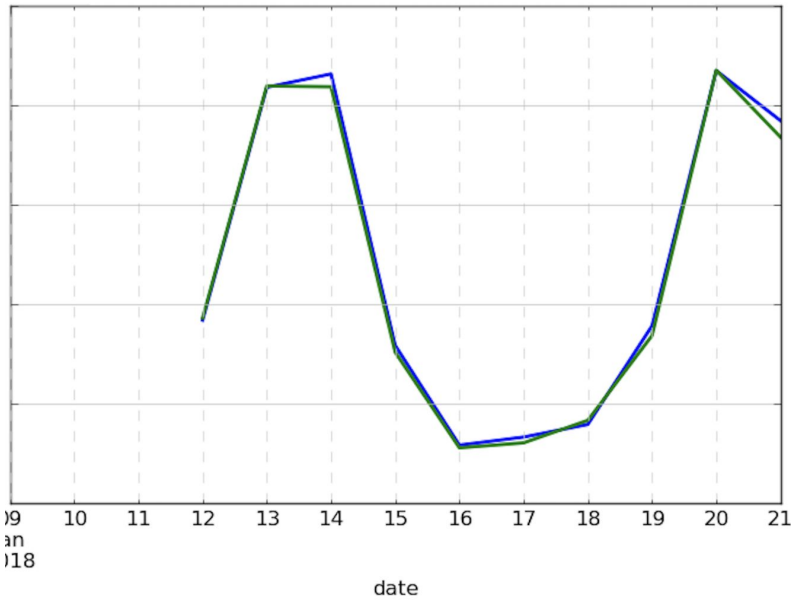

Figure 16: Two cohorts which are the same, but was a false positive due to a "split"

A third step is then added: for a day to be alerted, null hypotheses from *N* days prior must also be rejected. This greatly reduces the number of false positives, such that situations presented in Figures 15 and 16 do not occur.

6 Software Implementation

6.1 Overview

SMAD is driven by a nightly batch job which is set to automatically run after new data is loaded into the database. The driver is broken down into several steps for maximum modularization and maintainability:

1. Global metrics outlier detection
    a. Fetch data from database
    b. Data is cleaned, and metrics with insufficient data are automatically skipped
    c. Perform analysis
    d. Add results to email builder
2. Experiment metrics difference detection
    a. Fetch data from database
    b. Data is cleaned, and metrics with insufficient data are automatically skipped
    c. Perform analysis
    d. Add results to email builder
3. Generate plots
    a. Generate plots for outliers in global metrics

    b. Generate plots for differences in experiment metrics
    c. Add plots to email builder
4. Send out email alert to addresses, such as product managers and experiment owners
    a. Generate email with email builder
    b. Send out email
5. Save outliers and differences found to their respective tables
6. Send out separate email for any non-fatal exceptions caught to SLAM team members

6.2 Metric Analysis

For analysis of global metrics, the Shewhart weighted sum EWMA model with intelligent date indexing is implemented, as described in Section 4.5. For analysis of categorical experiment metrics, the G-test with William's correction from Section 5.1 is used following the 2x2 contingency table formulas found in [5]. Welch's *t*-test, described in Section 5.2, is implemented for discovering differences in continuous experiment metrics. The two additional steps for false positive rate control from Section 5.3 are also used for experiment metrics difference detection.

6.3 Building the Email

SMAD's email builder is implemented following the builder design pattern popularized by [3]. To offer a better user experience, special CSS styling is added to the HTML email to enable hover popups for displaying the outlier and difference plots. Samples of SMAD's email are shown in Figures 17 and 18.

Figures 17, 18 removed for confidentiality purposes.

6.4 Performance Optimizations

Since so many operations are performed by SMAD, parallelism was introduced into the application where applicable to improve performance. Steps 1abc, 2abc, 3ab, and 5 from Section 6.1 are implemented with customizable parallelization levels.

7 Discussion and Conclusion

This report summarizes my intern project while on Yelp's SLAM team. SMAD was built to automatically detect and alert on system failures characterized by outliers in global metrics, or statistical differences in the cohorts of experiment metrics. The data ingested by SMAD was described and several methods of detecting outliers in univariate, streaming data were proposed and compared for analyzing global data. This report describes how statistical tests were applied to

experiment metrics to find differences, and also provides an overview of how SMAD was implemented in code.

SMAD was written to be easily configured and modularized such that detection methods can be easily modified or completely swapped. A next step would be to manually tune some parameters to achieve better results. A Spark driver was also written to perform random search hyperparameter optimization, which should occasionally (time gap of months) be run to achieve better performance in outlier detection in global metrics. It is also recommended that improvements to the predictions in the Shewhart weighted sum EWMA model be made, using knowledge that certain days through the year are special and recurring, Christmas for example.

Acknowledgment

I would like to thank the following people without whom this project and report would not have been possible. In alphabetical order:

- Dhruv Saksena, for his mentorship, code reviews, coining the project, and help with Pandas
- Faizan Shabbir, for helping me with Spark and chatting with me when I hit roadblocks
- Michael Woods, for explaining parametric versus nonparametric tests to me
- Mirza Basim Baig, for his mentorship, code reviews, and helping me grow non-technically
- Sebastien Couvidat, for telling me about multiple comparison corrections and suggesting different methods to detect outliers in time series
- Yuhan Luo, for her code reviews and supplying metrics data

References

[1] C. M. Borror, D. C. Montgomery and G. C. Runger, "Robustness of the EWMA Control Chart to Non-Normality," *Journal of Quality Technology,* vol. 31, no. 3, p. 309–316, 1999.

[2] D. C. Montgomery, Introduction to Statistical Quality Control, 5th ed., John Wiley & Sons, 2007.

[3] E. Gamma, R. Helm, R. Johnson and J. Vlissides, Design Patterns: Elements of Reusable Object-Oriented Software, Addison-Wesley, 1994.

[4] H. Raza, G. Prasad and Y. Li, "EWMA Based Two-Stage Dataset Shift-Detection in Non-stationary Environments," 7 February 2017. [Online]. Available: https://hal.inria.fr/hal-01459655/document. [Accessed 17 April 2018].

[5] InfluentialPoints, "The G likelihood-ratio test," [Online]. Available: http://influentialpoints.com/Training/g-likelihood_ratio_test.htm. [Accessed 17 April 2018].

[6] J. H. McDonald, "G–test of goodness-of-fit," in *Handbook of Biological Statistics (3rd ed.)*, Baltimore, Sparky House Publishing, 2014, p. 53–58.

[7] J. W. Tukey, Exploratory Data Analysis, Addison-Wesley, 1977.

[8] R. R. Sokal and J. F. Rohlf, Biometry: The Principles and Practices of Statistics in Biological Research Third (3rd) Edition, New York: W. H. Freeman and Company, 1995.

[9] T. Finch, "Incremental calculation of weighted mean and variance," February 2009. [Online]. Available: http://people.ds.cam.ac.uk/fanf2/hermes/doc/antiforgery/stats.pdf. [Accessed 17 April 2018].

[10] Y. Benjamini, A. M. Krieger and D. Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika,* vol. 93, no. 3, p. 491–507, 2006.